DOCUMENT RESUME

ED 325 491 TM 015 708

AUTHOR Gershon, Richard C.
TITLE Rasch-Model Procedures Used To Build the JOCRF Vocabulary Item Bank. Technical Report 1990-3.
INSTITUTION Johnson O'Connor Research Foundation, Chicago, IL. Human Engineering Lab.
PUB DATE Sep 90
NOTE 60p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Cognitive Processes; Database Design; *Databases; Difficulty Level; *Item Banks; Item Response Theory; Statistical Analysis; Test Construction; *Test Items; *Vocabulary Development
IDENTIFIERS *Rasch Model; *Word Banks

ABSTRACT

In an effort to improve the ways in which words are learned, the Johnson O'Connor Research Foundation (JOCRF) is attempting to determine the difficulty level of all non-technical words in the English language. This item banking project entails: (1) identifying words that should be calibrated; (2) writing a test item for each word; (3) testing the item in public schools and private schools; and (4) calculating a series of statistics to assess the relative difficulty of a word and place it on the JOCRF's Vocabulary Scale. The vocabulary data base is composed of five data bases (ITEMS, USED, DISCUSS, STATS, and ALLSTATS) that are related to each other via various "key" fields. This report outlines the Rasch model statistical procedures used to determine the difficulty of a word. The rationale for using the Rasch model and a description of actual use of the statistical procedures are provided. The data base structure that the JOCRF uses to store the large quantities of statistical and verbal data generated by the project is also described. Eight appendices are provided containing numerous figures and tables that supplement the text. A 36-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED325491

RASCH-MODEL PROCEDURES USED TO BUILD THE JOCRF VOCABULARY ITEM BANK

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROBERT KYLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Richard C. Gershon

JOHNSON O'CONNOR RESEARCH FOUNDATION, INC.

Technical Report 1990-3

September, 1990

TM015908

COPYRIGHT © 1990 BY JOHNSON O'CONNOR RESEARCH FOUNDATION, INC.
ALL RIGHTS RESERVED

Rasch-Model Procedures Used to Build the JOCRF Vocabulary Item Bank

Richard C. Gershon

Abstract

The Johnson O'Connor Research Foundation is actively pursuing research to improve the ways in which one learns words. In this regard the Foundation is attempting to determine the difficulty level of all nontechnical words in the English language.

The item banking project entails several operations: the identification of words that should be calibrated, the writing of a test item for each word, the testing of that item in public and private schools, and the calculation of a series of statistics that assess the relative difficulty of a word and place that difficulty on the Foundation's Vocabulary Scale.

This report outlines the Rasch-model statistical procedures that the Foundation uses to determine the difficulty of a word. The report gives both the rationale for using the Rasch model and a relatively nontechnical description of how the statistical procedures are actually used. The report also describes the database structure that the Foundation uses to store the large quantities of statistical and verbal data that are generated by the vocabulary project.

Contents

Introduction 1

Equating Items to the Johnson O'Connor Vocabulary Scale 2

The Vocabulary Database 6

Appendices

A - The Rasch Model and the Item Characteristic Curve 8
B - Constructing a Rasch Item Bank for the Johnson O'Connor Vocabulary Tests (by Richard Smith) 11
C - Linking Constants Obtained Using Various Equating Strategies . . 16
D - Linking Structure, Items, and Anchor Values 18
E - Sample BIGSCALE Command File 26
F - Sample BIGSCALE Output 30
G - Database Structure 39
H - Test Series Contained in Each Database 48

References 51

Acknowledgments

Research into arranging words in their order of difficulty began in the early 1930s when Johnson O'Connor emphasized the unique importance of vocabulary knowledge. In more recent years many persons have contributed to this effort. Foundation President George Wyatt and vocabulary program coordinator Steve Aldrich work with item writer William Shapiro to write the test items and the subsequent *Wordbook* discussions. In the past, Richard Bowker, Gary Supanich, and Bruce Ingram wrote test items and *Wordbook* discussions. Not to be forgotten is Mary Lou McCarty of the Houston office, who locates hundreds of schools each year willing to participate in the calibration process. I would also like to thank George Wyatt, Thomas McAveeney, Robert Kyle, David Schroeder, and Steve Aldrich, who offered helpful comments on this manuscript.

Introduction

The Johnson O'Connor Research Foundation has had a commitment to the study of aptitudes and vocabulary acquisition since its founding in 1922. With regard to vocabulary, the Foundation maintains several ongoing programs: testing, education, and research.

In terms of testing, examinees who take the Foundation's testing battery currently are administered Worksample 690, the latest in an extensive series of vocabulary tests, consisting of a total of 225 items, divided among three overlapping forms (easy, intermediate, difficult). A given examinee takes a short placement test (Worksample 695) and then the appropriate form of Worksample 690. The resulting raw score is converted to what is referred to as a Vocabulary Scale Score (VSS). The VSS value is the raw score that would be obtained on Worksample 690 if all 225 items were administered (Statistical Bulletin 1980-33). The scale defined by VSS values is used by the Foundation as a common scale against which all vocabulary tests and vocabulary items can be referenced. The VSS scale also allows the vocabulary abilities of persons to be placed on a single continuum for comparison and norming purposes. Previously, examinees took all 225 items on Worksample 690. The use of the three overlapping forms eases the burden on low-vocabulary examinees, who took many items beyond their ability level on the longer test. The shorter forms also ease the burden previously placed on high-vocabulary examinees, who took many easy items that did not help to discriminate their vocabulary ability (Statistical Bulletin 1980-33).

In order to place future items on the VSS scale, Worksample 705-1 was designed to contain 75 of the Worksample 690 items, which were referred to as the "equating" items, and a group of new, easier items, which have come to be known as the Foundation's principal set of 60 "linking" items. Worksample 705-1 was administered to 212 junior and senior high school students in the spring of 1983. The difficulties of the 60 linking items were calculated and placed on the VSS scale. The 60 linking items could then be used to "link" future experimental items to the VSS scale. These linking items were administered along with experimental items on subsequent high school testing series to link the new items to the Foundation's VSS scale (see Appendix D).

In the field of education, the Foundation publishes a vocabulary building series known as *Wordbooks* (Bowker, 1979a, 1983). Each *Wordbook* contains teaching exercises for a group of 180 vocabulary words that fall within a narrow difficulty range, defined by the words' VSS values. One goal of the Foundation's vocabulary

research program is to determine the VSS values of all the nontechnical words in the English language. These words will eventually be used in additional *Wordbooks*. By knowing the relative difficulty of all English words, the Foundation can suggest the words that are the most appropriate for teaching to a group of a given vocabulary ability.

Equating Items to the Johnson O'Connor Vocabulary Scale

In terms of research, the Foundation now uses the Rasch measurement model to determine item difficulties and "equate" those difficulties with the VSS Scale (see Appendix B for a description of the Rasch model). There are numerous methods that can be used to equate tests and items within tests using the Rasch model. In this regard, Richard Gershon, Research Department Research Assistant, and David Schroeder, Research Manager, have conducted research that showed that for Foundation vocabulary items, there are no substantial differences between the commonly used equating strategies (see Appendix C; Gershon & Schroeder, 1987). This research led us to the conclusion that "item anchoring" is the best method to use because it is the most time-efficient and because it produces item files and printouts with the equated item statistics. In brief, item anchoring is the process by which the values of the linking items are fixed at their VSS values in the analyses of the difficulties of the experimental items, so that no further equating is necessary (see Schultz, 1988, and Kelderman, 1986).

The Foundation now uses the Rasch-model software for personal computers called BIGSCALE (Wright, Linacre, & Schultz, 1989) to calculate item statistics. Prior to the Worksample 741 test series, administered in 1989, a similar program called MSCALE (Wright, Congdon, & Rossner, 1987) was used. I will detail the procedure that we follow to analyze data with BIGSCALE:

1. A raw data file that contains the answers chosen by examinees for a single vocabulary test form is constructed. Typically this form will consist of 36 linking items (a reduced set of the original 60 items) and 74 experimental items.

2. The linking items' difficulties are placed in a BIGSCALE-compatible "anchor file" with their predetermined values in "logit" units. Logits are the units that the Rasch model uses to express person abilities and item difficulties. Logit scales are desirable because they quantify the given variable on a linear, interval-level scale (see Appendix A). As noted earlier, the original anchor values for linking items were established when Worksample 690 was equated with the 705-1 linking items, and the

use of these anchor values ensures that all new items tested by the Foundation are placed on a common scale (see Appendix D for a complete overview of the linking structure). Beginning with Worksample 738, the set of linking items was reduced from 60 items to 36. Two additional sets of linking items were constructed for Worksample 741, one set for use in early primary grades and the other for use with college students. The list of the three linking item sets and their anchor values can be found in Appendix D.

3. The data set is analyzed with BIGSCALE (Wright, Linacre, & Schultz, 1989). The printouts and item statistic files are generated so that the item measure is automatically placed on the Foundation scale without further equating.[1] A sample program is shown in Appendix E. A sample printout of the results is in Appendix F.

4. The difficulty value, or "logit measure," for each item (and person) corresponds to the vocabulary ability level at which an examinee has a 50% probability of getting the item correct. Logit measures can be converted to VSS units by the following linear transformation:

(MEASURE X 26.78) + 128.40

Since the *Wordbook* program uses items at the *80%* level, however, the formula appropriate for conversion to 80% is as follows:

(MEASURE X 26.78) + 165.40

The first formula was derived by regressing the VSS values of a group of Worksample 690 items on the logit measures obtained for those same items. This method determined that each logit represents 26.78 VSS units. In the Rasch model the point representing the 80% chance of getting an item correct is always 1.38 logits from the 50% point. Multiplying this by the logit size given in the first equation results in the second equation.

Although we have used these linear formulas, it may be the case that a single linear transformation is not accurate across the entire VSS range. Further research should be conducted to determine whether this is the case and whether additional formulas need to be derived.

---

[1]BIGSCALE also allows for the one-step computation of person ability estimates on the Foundation VSS scale expressed in Rasch logits. These estimates are computed at the same time that the items are calibrated and are provided along with person fit statistics.

5. After VSS values are obtained, it is necessary to determine the quality (validity) of each item. A good-quality item is one for which people below a given level consistently get the item wrong, while people above the level consistently get it right. It is possible, for example, that a poor-quality item was guessed correctly by an extremely large number of low-vocabulary people, or that one of the misleads was so attractive that people who would ordinarily have known the word answered that item incorrectly. Fortunately, BIGSCALE provides several statistics that make it fairly easy to determine the quality of the item. The most important of these statistics are called INFIT and OUTFIT. Fit statistics serve much the same function in item response theory as the item-total correlation in classical test theory.[2] They provide a measure of how well an item agrees with the total test score.

In general, a fit value near zero indicates an average degree of agreement between the item and the total test score. A *negative* value indicates a *better* than average degree of agreement between the item and the total test score. The lower the negative value, the greater the level of agreement. A *positive* value indicates a *poorer* than average degree of agreement between the item and the total test score. The higher the value, the lower the level of agreement.

In other words, an item has a negative fit if persons with word knowledge better than the item difficulty almost always get the item correct, and persons with word knowledge below that of the item difficulty almost always get the item wrong. For the majority of items the fit value is near zero (plus or minus two), indicating that the more-competent persons *usually* answered the item correctly, and the less-competent persons *usually* answered the item incorrectly. Positive fit statistics indicate that at least one of the two conditions was not met--that is, a relatively large number of higher-vocabulary persons answered the item incorrectly and/or a relatively large number of lower-vocabulary persons answered the item correctly.

5a. Item INFIT is roughly equivalent to the ability of an item to accurately discriminate in the vicinity of its difficulty level. In VSS terms, it means that for an

---

[2]Classical test theory was first presented by Charles Spearman. He posited that test scores were actually the sum of two components: the person's "true" score plus an error component. Using classical methods, a person's measure on a given variable is solely determined by his total test score. The problem with this approach was that two good-quality tests could be constructed, but if one consisted of easier items than the other, the results of the tests would differ and would not be directly comparable. In addition, the estimated difficulty of the test was directly related to the ability of the sample taking the test, with no direct method of relating the same test to a more- or less-able sample (Mislevy, 1990).

item of VSS 100, only people with a VSS score of 100 or more should answer the item correctly. In some situations, such as in the use of linking items used to equate populations, highly negative INFIT values would be undesirable (personal communication, Benjamin Wright, April 26, 1990). However, for the *Wordbook* testing program, negative INFIT values are probably just fine.[3] (As noted, items that are used for test linking should probably *not* have INFIT values greater than + 2.0.) Many people use + 2.0 as their cutoff for item INFIT, but since our samples are usually relatively large (i.e., 400-500 students), we reject only items with INFIT values greater than +4.0. Items with INFIT values of this magnitude should be rewritten and retested.

5b. Item OUTFIT is similar to INFIT, but it is more sensitive to unexpected correct responses by low-vocabulary examinees and to incorrect responses by high-vocabulary examinees who are far above the VSS level of the item. Items with OUTFIT values greater than +4.0 should be rewritten and retested. Oftentimes a large OUTFIT value is obtained when a large number of low-ability people guess an item correctly. This may be due to such things as the item being too difficult for the sample population (see also Point 6), or to the misleads being so unattractive that they were never chosen. (Items that are to be used for linking also should not have an OUTFIT value more than 2.0 units away from the INFIT value; this may indicate guessing or order effects[4] for the item.)

5c. The Mean Square statistics provide additional measures of the quality of an item. They refer to the ratio of the item variance that actually occurred to that which was expected, given the item measure that was obtained and the ability levels of the people in the sample. The maximum value of Mean Square that should be acceptable for INFIT is 1.2, meaning that a maximum of 20% of the item variance is unexplained by the model. BIGSCALE also produces a value for Mean Square OUTFIT. Based on the results of the Worksample 741 data analysis, it would appear to be reasonable to select a maximum acceptable Mean Square OUTFIT value of 1.4.

6. The final issue that must be addressed is whether the item was given to the correct population. If the item was much too easy or much too difficult for the

---

[3]Negative fit values are satisfactory except in the case where the ability of the sample differs substantially from the difficulty of the item (personal communication, Benjamin Wright, May 1989). Because of this, the bad sample criterion should probably be made more strict for items with high negative INFIT.

[4]An order effect occurs when an item's difficulty changes depending on its position in a test.

persons who took it, the difficulty estimate for that word will not be accurate. Items that are eliminated for either of these reasons are not necessarily "bad" items, but they must be retested with an appropriate population. The use of estimates of word difficulty, such as those in *The Living Word Vocabulary* (a word list that includes over 40,000 words and gives a percentage score for knowledge by persons of varying grade levels; Dale & O'Rourke, 1981), limits the number of items that must be retested.

The acceptable difficulty range for items depends on the sample population. Any item that is more than plus or minus 2.5 logits from the mean person measure for that form should be retested. The mean person measure for a form can be found in Table 20 of the BIGSCALE printout for that form (see Appendix F).

I have described several special considerations to be used in selecting linking items (see Points 5a and 5b). I would further suggest that no short explanation would sufficiently cover all the contingencies and issues that may arise in selecting linking items. For example, linking items should adhere to stricter selection criteria regarding fit statistics than should other vocabulary items. Otherwise, an item that appears to be of good quality when administered to low-ability examinees may end up being of poor quality when administered to high-ability examinees. When this occurs, the quality of the linking deteriorates, and new experimental items that are presumed to be correctly linked to the Vocabulary Scale will have inaccurate difficulty values. Therefore, linking items should probably not be selected without the aid of someone well-versed in the Rasch model. Since 1983, the Research Department has received consultation from Dr. Benjamin Wright regarding our choices of linking items, equating procedures, and Rasch-model software.

The Vocabulary Database

The vocabulary database is composed of five databases that are related to one another by means of various "key" fields. (A relation in computer software refers to the capacity to look something up in one file and automatically be able to find related information in other files.) The structures of the five databases can be found in Appendix G. The test series covered by the databases can be found in Appendix H. The following is a description of the databases and how they can be used:

ITEMS — This database is a collection of all the items that have been calibrated, whose test words are available for use in future *Wordbooks*.

USED	This database contains all the vocabulary items ever tested by the Foundation.
DISCUSS	This database contains *Wordbook* discussions, pretest items, exercises, and review test items for future use in *Wordbooks*. Many of the Worksample 705 and 722 words have completed records in DISCUSS. (Technical note: the memo fields from DISCUSS.DBF are actually maintained in DISCUSS.DBT. This file should not be erased.)
STATS	This database lists all the item statistics that have been computed for Foundation items. This database should be maintained by the Research Department as a statistical archive. Although item difficulty values are contained in this file, the VSS value for each item can also be found in ITEMS. STATS contains items from test series where the only statistic listed may be the VSS value.
ALLSTATS	A second database similar to STATS is ALLSTATS, which contains items for which we have Rasch-model measurement statistics. A recent list of items with acceptable statistics was published by the Foundation in 1988 (Technical Report 1988-3).

Appendix A

The Rasch Model and the Item Characteristic Curve

The *Teacher's Manual* of the *Wordbook* series outlines the method originally used by Richard Bowker of the Foundation for determining item difficulty (Bowker, 1979b, pp. 4-6). In brief, a subset of the Worksample 690 items was administered along with the experimental items. Bowker then graphed the proportion of persons who answered the item correctly for various score ranges on the Worksample 690 subset. The difficulty of the item was defined as the Vocabulary Scale Score where 80% of the examinees answered the item correctly. When *Wordbooks 7* and *8* (Bowker, 1983) were added to the series, Bowker began to use the Rasch measurement program BICAL (Bowker, 1982; Wright, Mead, & Bell, 1980). His use of BICAL was still graphical in nature, however, as he relied on examination of the BICAL charts to determine the item's difficulty.

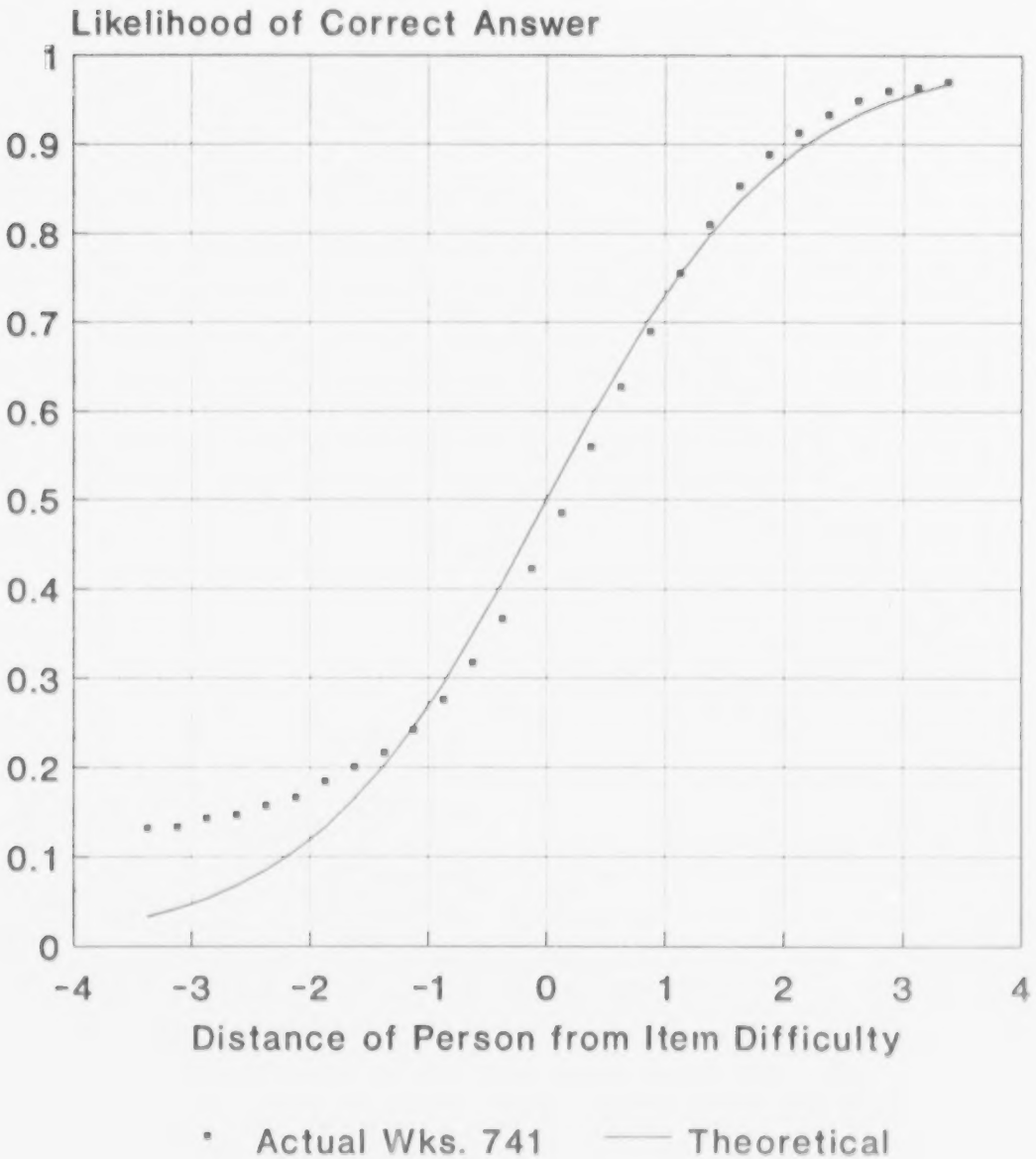
Beginning with the 705 series, the Foundation began to use the Rasch model statistics themselves (see Appendix B). The Rasch model employs a theoretical curve for each item that relates the percentage of persons who answer an item correctly to the log-linear difference between their ability and the difficulty of the item. The "Theoretical" curve presented in Figure 1 shows the likelihood of a person answering an item correctly given the distance of that person's vocabulary knowledge from the item difficulty (in logits). You will note that when the distance is zero (the person's ability is the same as the item's difficulty), the person has a 50% likelihood of answering that item correctly. As described in the equating section of this report, however, the Foundation defines the difficulty of an item as being the point where 80% of the persons answer the item correctly. In logit terms this is equivalent to adding 1.3863 logits to the difficulty of the item.

Figure 1 also shows the actual percentages obtained using Worksample 741 compared to the theoretical values predicted by the Rasch model. The "Actual Wks. 741" line indicates the average results obtained for Worksample 741 across 22,000 students taking 110 items each. The "Theoretical" line indicates the percentages suggested by the Rasch model. As you can see, there appears to be satisfactory agreement between the two methods at all levels except for extremely low-ability persons, for whom chance becomes an issue (theoretical values below 20%).

Given the close agreement between the Rasch model and the observed likelihood function for vocabulary items, a clear case can be made for using the Rasch model (also see Appendix B). The Rasch model allows a single pass to be made of the data,

Figure 1

Comparison of Actual versus Theoretical Item Characteristic Curve of Wks. 741



9

using a computer program such as BIGSCALE (Wright, Linacre, & Schultz, 1989) to simultaneously compute the person abilities and the item difficulties. It should be noted that this approach uses all the data to estimate item difficulty, whereas the graphical approach outlined by Bowker makes no use of data away from the determined difficulty level. The Rasch approach also uses a smooth curve that corresponds closely to the actual probabilities, using a linear (interval-level) scale (Wright, 1977a).

Appendix B

Constructing a Rasch Item Bank for the Johnson O'Connor Vocabulary Tests

Richard Smith

Recently, there has been a great deal of interest in applying latent trait theory to test development research. Latent trait models are useful because they provide a way of analyzing and interpreting responses to items independently of the ability of the sample used. Rasch recognized that objective measurement requires person measures that do not reflect the particulars of the items used. Furthermore, the ordering of the items that define a variable should be independent of the persons measured (Rasch, 1960, 1961, 1966a, 1966b, 1977; Wright, 1968, 1977b).

The primary task of psychological measurement is to ascribe meaning to scores in such a way as to establish a joint order of persons and items along a single common linear scale. To measure and understand individuals, we must construct person-item interactions that provide insight into the degree to which a person possesses the aptitude.

Tentatively we may consider what happens to the probability of a person succeeding on a test item. The Rasch model has only one ability parameter B for each person and only one difficulty parameter D for each item i. The probability of a right answer is determined by the difference between ability and difficulty (B-D) expressed as a ratio of natural logs. Persons with more ability should always have a greater probability of answering any item correctly than persons with less ability. Easy items should always be answered correctly more often by everybody than hard items. If the response person n gives to item i is expressed as $X_{ni}=1$ for a correct response, and $X_{ni}=0$ for an incorrect response, then the Rasch model for measuring persons and calibrating items becomes:

$$P[X_{ni}=1]=\frac{e^{(B-D)}}{1+e^{(B-D)}}$$

The Rasch model is the only latent trait model where the unweighted sum of right answers given by a person is a sufficient statistic for the person's ability. This means that the conditional probability of the item responses of an examinee, given the person's raw score, is independent of the examinee's ability. Similarly, the unweighted sum of right answers given to an item will contain all the information necessary to calibrate that item along the variable.

The uniqueness of the Rasch model focuses on the concept of specific objectivity. This is formalized by Wright as test-free person measurement and sample-free item calibration. Objectivity involves logical order, parameter separation, and estimation efficiency. The property of logical ordering implies that for any person, the probability of success is greater for an easy item than for a hard one; for any item, an able person has a greater probability of success than an unable one. The ordering of every person and every item along a single common variable allows objective comparisons among persons and items.

A basic requirement for Rasch measurement is that the variable being measured is unidimensional, so that a single score is meaningful and useful. This has many practical implications. It means that a person's ability is all that is needed to predict his performance on a set of test items. It is not necessary to know anything about what group the person belongs to or what year the person took the test. It also means that all persons moving in the same direction along the line of the variable must pass through the same points in the same order. We can make probability statements about any person encountering any item, based on an estimated ability and difficulty. The items that are ordered along the line provide the operational definition of the variable. The relative positions of items on the line are determined by the performances of persons on those items.

Although the raw score is a sufficient statistic, it must be transformed into a linear and objective measure of the person's position on the variable. The logistic transformation stretches the score at the extremes so that the resulting logits are linear in the ability implied by the score. Linearity means that an increase of one unit represents the same increment in ability at any point along the scale. Although the raw score is specific to the test, the logit measure is general on the variable.

In summary, the structure of the Rasch model in which parameters enter linearly without interactions makes the complete separation of the model's parameters possible. As a result, the likelihood equations can be written so that it is possible to derive conditional estimation equations for person abilities and item difficulties that are completely independent of each other.

Separable person and item parameters permit the calculation of sufficient statistics that are simple counts. These are the number of right answers for each person and the number of successful persons for each item. Since all information about the abilities of the persons is contained in their raw scores, the estimation equations for the item difficulties can be expressed in terms of the unknown difficulties and the observable person scores. These sufficient statistics correspond to the greatest data reduction that can be achieved while still defining the likelihood.

The essential aspects of specific objectivity cannot be separated from each other. A unique ordering of persons and items to be inferred from the data is crucial. This inference requires a probabilistic Rasch measurement process which has separable parameters and hence sufficient statistics. These properties and their psychometric implications are described in Rasch (1960, 1968), Andersen (1970, 1973, 1977), and Wright (1968, 1977a, 1977b, 1985).

Item Bank Building

An item bank is a collection of carefully calibrated test items that define a variable. It is a continually evolving measurement system in which the systematic assessment of educational achievement or acquired knowledge is a permanent activity. This section describes several features of item banks, including the motivation for banking techniques. The primary incentives that justify the effort required to establish and maintain a Rasch-based item bank are meaning and convenience. Meaning comes from the careful delineation, over a broad range of application, of the variable that the items in the bank are designed to measure. Convenience comes because a calibrated bank makes it easy to construct and equate new forms for a variety of purposes.

A clear, unequivocal, and objective definition of the variable to be measured is fundamental to the success of any measurement task. For the Foundation, it must be possible to imagine that the particular knowledge or aptitude of each examinee can be described quantitatively on a scale. Apart from the quantitative attributes of such a scale, any meaning that is to be attached to it must come from the items that are used to observe it. It is only through the placement of items along the continuum according to their relative difficulty that we can understand what it means for an examinee to be at a particular location along the continuum. Once the items are located, the bank is, in principle, built.

An essential psychometric quality of Rasch item banking is that when items are calibrated onto a common variable, each item represents a position on the variable that is also represented by other items of comparable difficulty. This makes it possible to infer an examinee's mastery with respect to the basic variable that the items share, regardless of which items are administered or whom else has been tested (Wright & Bell, 1984). Each person's position on the variable places that person among whomever else has ever taken any set of items from the bank. For example, examinees will receive scores that are commensurate with their current knowledge of English vocabulary, irrespective of which items from the bank are used to assess their knowledge.

The most fundamental part of developing an item bank is to objectively define the variable and to locate items along a line according to their relative difficulty. When a

variable is mapped in terms of its items, then standards can be established and meaning attached to being at a particular point along the variable.

After all items are calibrated onto a common linear scale, any subset drawn from the bank will be automatically equated to the bank and to any other possible set of bank items. This is achieved without any further testing. As a result, it is simple to equate tests from year to year or to equate multiple forms given on the same occasion. Scores a person makes from time to time are directly comparable and the rate of progress apparent. Choppin (1978) provides a comprehensive examination of the conceptual issues and psychometric implications of item banking and item calibration.

Since many persons do not follow our expectations of which items are easy and which are hard, we can apply Rasch's probabilistic model to impose an orderly response process on the data (Wright & Bell, 1984). In order to have a common basis for describing progress, there should be agreement among researchers and examinees as to which items are hard and which are easy.

Several steps are necessary to build and maintain an item bank:

1) Designing test forms.
2) Calibrating test forms.
3) Analyzing fit.
4) Linking pairs of forms.
5) Calibrating forms on the bank.
6) Analyzing link fit.
7) Controlling item quality.
8) Monitoring and updating the bank.

First, items are written and distributed among test forms so that there is a network of common items that is practical to the testing situation. Forms are designed and administered. The process of calibrating sample-free item difficulties is performed under the expectation that these data can be used to approximate additive conjoint measurement (Brogden, 1977). The calibration of items that is sample-free and the measurement of persons that is test-free are the precious ingredients that supply the natural fuel needed for the development of a successful item bank measuring system.

The computer program BICAL (Wright, Mead, & Bell, 1980) [Note: more recently, MSCALE (Wright, Congdon, & Rossner, 1987) and BIGSCALE (Wright, Linacre, & Schultz, 1989)] was used to derive estimates of person abilities and item difficulties and to test the fit of items within each of the vocabulary tests. These item difficulties are invariant with respect to the ability of the calibrating sample; however, they are defined by the center

of the items in the specific examination. An item will appear to have a different difficulty for each test in which it appears, so we must adjust all difficulties on all exams so that they are positioned relative to one common origin. This requires linking together all relevant tests by calculating translation constants that shift these items to a common bank reference scale. The technique implies that if test X and test Y share a common set of K items, called the link items, the difficulty scale of test Y is adjusted to the scale of test X. Therefore, the link between two tests is estimated by the difference between the difficulties of any item calibrated in both exams. Common items between any pair of forms provide a direct estimate of the relation of the two forms. If the common items and the other items in both tests fit the Rasch model and are calibrated on the same latent variable, this method yields a pool of calibrated items whose estimated difficulties are on a common scale with a common linear metric.

After the bank has been constructed, it will need constant monitoring to verify that no item has lost its effectiveness. If some items are becoming too familiar or have been used too much, these items can be simply removed from the bank without disturbing the other items. An item difficulty is estimated every time an item is administered. When this estimate is statistically different from the item's bank difficulty, then thought must be given to what may have caused this change and how to resolve it. Whenever a new exam is given that uses items from the bank, it will be necessary to calibrate the new form and determine the appropriate translation constant to link the new exam to the existing bank through the reused items. New items must be introduced into the bank in the same way the original items were established when the bank was created.

Wright & Stone (1979, Chapters 5 and 6) describe procedures for calibrating tests and constructing item banks using the Rasch model. Further issues concerning the curricular implications of item banking and the psychometric basis of banking, along with computer programs and equations for accomplishing banking are presented in Wright & Bell (1984). Millman & Arter (1984) discuss the vast array of item bank features that allow them to operate effectively within diverse instructional and assessment environments.

Appendix C

Linking Constants Obtained Using Various Equating Strategies

Numerous equating strategies are suggested in the literature of item response theory. To select an equating strategy for the Foundation's vocabulary item banking project, five commonly used methods were carried out for three test forms: Worksample 705-2, Worksample 722-5, and Worksample 722-9.

Method 1 is the simplest. One simply averages the difficulties of the 60 linking items and computes the difference between that average and the average for the same items when administered on Worksample 705-1.

Method 2 employs a complex set of spreadsheet calculations to limit the set of linking items to those with standardized residuals below a particular level (see Wright & Stone, 1979, for a complete description). Method 2a gives the linking constant obtained when the residuals of the linking items were limited to a maximum value of 3. In other words, after the calculations are carried out, some of the linking items are discarded for the given form because their standardized residual values are greater than 3. The difficulties obtained on the remaining items are averaged, and the linking constant is computed as the difference between this average and the average of the same limited set of linking items administered on Worksample 705-1. Method 2b is similar to 2a except that the residual requirement is stricter and the retained linking items must obtain values less than 2. This results in an even smaller set of linking items being used.

Method 3 appears to be similar to Method 2 in that first the 46 best linking items were selected from the Worksample 705 test series using a standardized residuals analysis that included all the Worksample 705 test series forms. Although all 60 linking items were left on the test, the difficulty values of the 46 items were anchored in the MSCALE analysis of each form. While the previous two methods required MSCALE to be run, and then a linking constant to be added to the item difficulties obtained, item anchoring within MSCALE allows the printouts to include item difficulties already linked to the Foundation's Vocabulary Scale. It should be noted that none of the anchored items were deleted from the analyses.

Method 4 is the same as Method 3 except that the anchor items were selected from a spreadsheet residuals analysis of all the Worksample 722 forms.

Table 1 shows the results of the above linking and anchoring strategies when applied to each of three test forms. The values within the table are the effective linking constants for the various methods.

Table 1

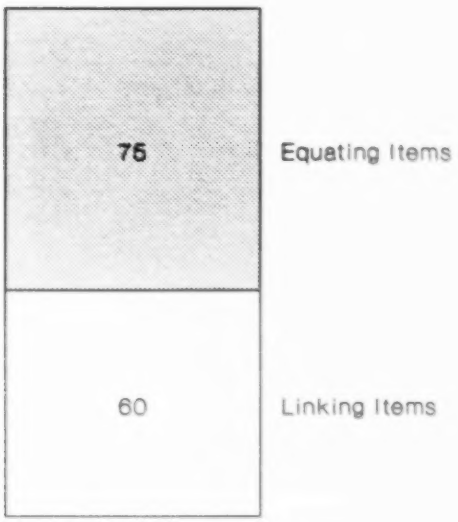
Anchoring Strategy	705-2	722-5	722-9
1) All 60 705-1 linking items	-3.336	-2.658	-2.250
2) Custom spreadsheet selection			
a) Standardized residuals < 3	-3.295	-2.69	-2.263
b) Standardized residuals < 2	-3.288	-2.67	-2.276
3) 46 preselected 705-1 links	-3.319	-2.668	-2.264
4) Best 32 links chosen from across all 722 forms	-3.346	-2.673	-2.294

Simple observation leads one to conclude that the differences between the above methods are negligible. This led us to conclude that item anchoring was the superior linking strategy because it is the most time-efficient and results in MSCALE outputs that already place all the items on the Johnson O'Connor vocabulary scale.

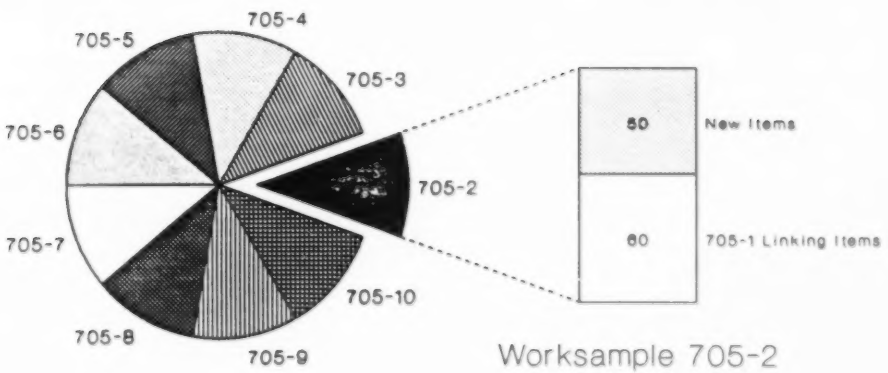
Appendix D

Linking Structure, Items, and Anchor Values

Worksample 705-1

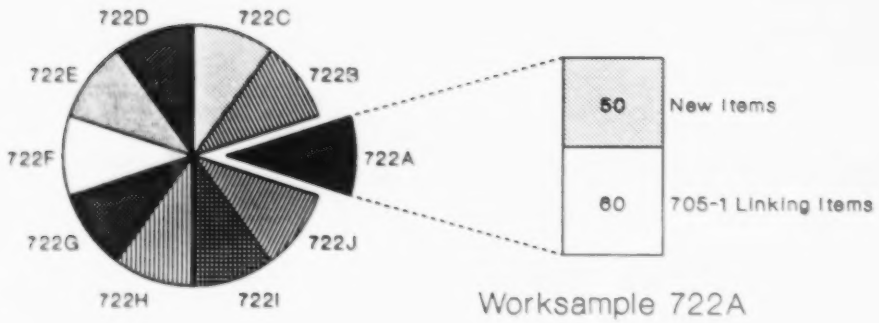


Worksample 705 Forms 2-10



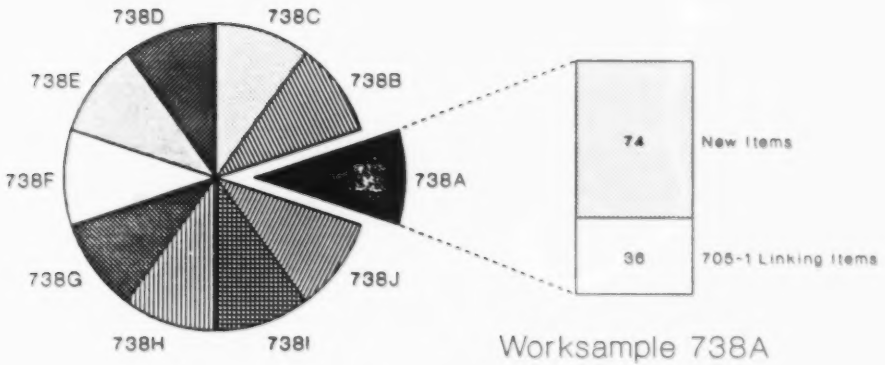
Worksample 705-2

Worksample 722 Forms A-J



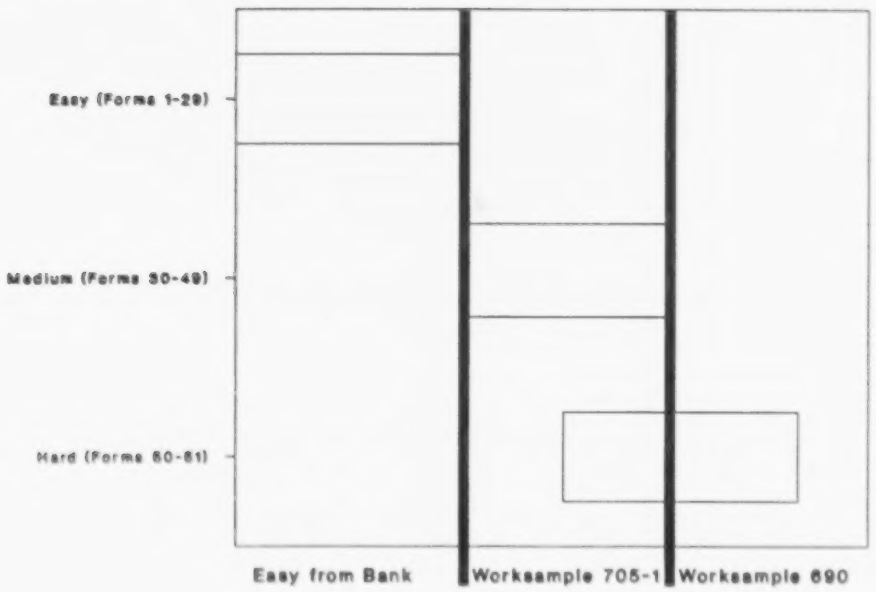
Worksample 722A

Worksample 738 Forms A-M
(Only Forms A-J shown)

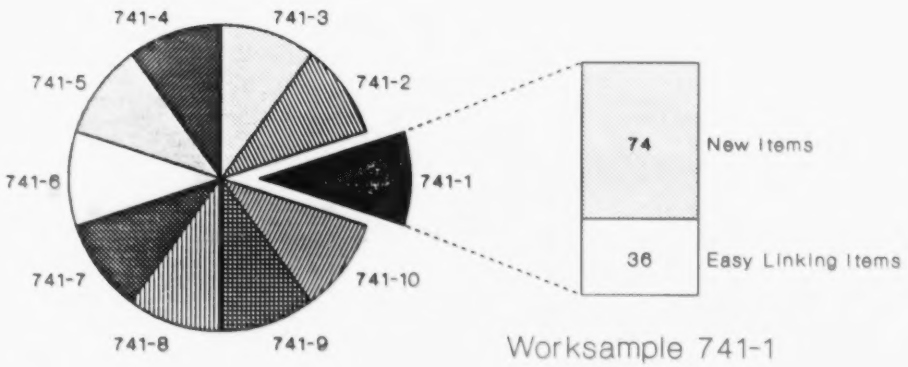


Worksample 738A

Origins of Wks. 741 Linking Items

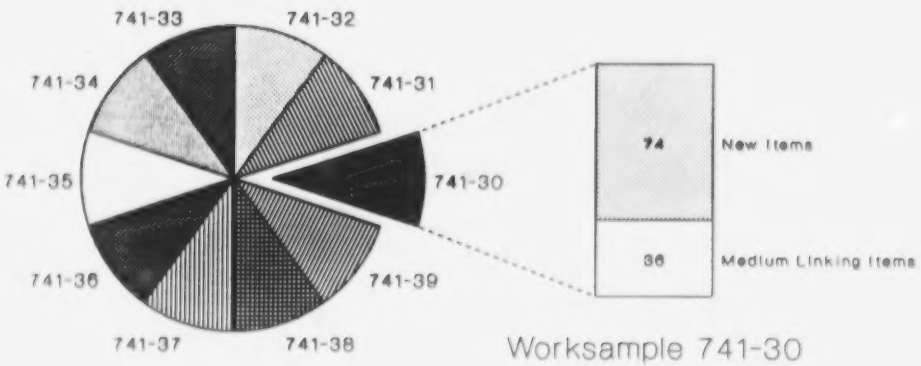


Worksample 741 Forms 1-29
(Only Forms 1-10 shown)



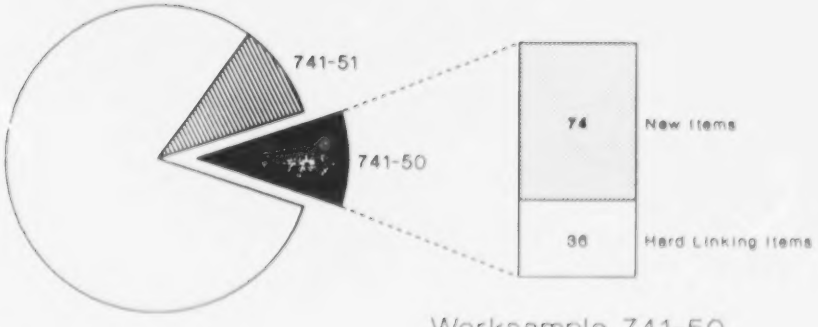
Worksample 741-1

Worksample 741 Forms 30-49
(Only Forms 30-39 shown)



Worksample 741-30

Worksample 741 Forms 50-51



Worksample 741-50

For the following lists, all anchor values are expressed in logits on the Foundation's Vocabulary Scale (at the 50% level).

Current Easy Linking Items and Anchor Values

TOPIC	-5.60
SHALLOW	-5.90
NURSED	-6.02
GAP	-4.56
COZY	-5.68
CABLE	-4.11
GUIDE	-5.22
NATURAL	-5.65
GNAWED	-4.97
JOURNEY	-5.35
COMPRESS	-4.45
GRAVEL	-5.09
SHRIVEL	-4.33
GRANTED	-4.30
FLOCK	-4.56
CRAM	-4.56
COMMOTION	-4.16
DECLARE	-4.23
WITHDRAW	-4.02
VALUE	-4.15
PRECISE	-3.75
APPROPRIATE	-3.77
GRIEF	-3.61
GLOBAL	-4.72
SLAY	-3.52
FRACTION	-3.59
BUREAU	-4.10
INFURIATE	-2.94
POURED	-3.86
BRISK	-2.85
UNIFORM	-3.68
EMPHASIZE	-2.83
UNSAVORY	-2.71
MUDDLE	-3.33
CONCEITED	-2.58
CURVATURE	-4.01

Current Intermediate Linking Items and Anchor Values

VANISH	-5.80
SHRIVEL	-4.33
ABSURD	-4.14
TASK	-4.78
APPROPRIATE	-3.77
ZANY	-4.59
SHRIEK	-4.53
COMMOTION	-4.16
HEX	-4.33
ASSAULT	-4.22
INTERNAL	-4.06
PRECISE	-3.75
WEARY	-3.89
GRIEF	-3.61
POSSESS	-3.94
SLAY	-3.52
RIGID	-3.54
DISMAL	-2.90
EMPHASIZE	-2.83
BARRICADE	-3.34
BRISK	-2.85
BADGER	-3.05
INFURIATE	-2.94
EXUBERANT	-2.83
DEVASTATE	-2.26
UNSAVORY	-2.71
CONCEITED	-2.58
BLEMISH	-2.48
INQUISITIVE	-2.16
PUTRID	-2.08
SERENE	-2.03
CLAMOR	-2.26
MEAGER	-2.34
ABHOR	-1.83
MONUMENTAL	-1.92
ACKNOWLEDGE	-2.03

Current Difficult Linking Items and Anchor Values

RIGID	-3.54
BARRICADE	-3.34
EXUBERANT	-2.83
CONCEITED	-3.56
BLEMISH	-2.48
MEAGER	-2.34
CLAMOR	-2.26
ACKNOWLEDGE	-2.03
MONUMENTAL	-1.92
ABHOR	-1.83
SOUVENIR	-3.94
AGHAST	-2.56
REPLICA	-3.94
AGITATED	-4.63
RESPONSIVE	-3.26
DETESTED	-3.31
TERMINATION	-5.52
INCISION	-4.35
PROLONG	-4.79
FRACTURE	-2.92
DELUSIONS	-3.56
STIMULATED	-3.94
VERBOSE	-2.20
REPULSIVE	-3.20
DETERIORATED	-1.16
BESEECHES	-0.25
ULTIMATUM	-2.09
LEISURELY	-2.06
RIGOR	-1.84
SANCTITY	-1.32
SCRUPULOUS	-0.06
SUBORDINATE	-1.27
CAPRICE	-0.41
ASSUAGING	0.61
EFFRONTERY	0.49
ERUDITE	1.78

Previously Used Linking Items and Their Anchor Values
(used with Worksamples 705 & 722)

ABSURD	-4.14	SACRED	-3.14
ABUNDANT	-3.61	SLAY	-3.52
APPROPRIATE	-3.77	TRIBUTE	-3.21
ASSAULT	-4.22	ZANY	-4.59
COMMOTION	-4.16	EMPHASIZE	-2.83
DISPUTE	-3.61	BADGER	-3.05
VANISH	-5.79	FEEBLE	-3.56
GRIEF	-3.61	NONCHALANT	-2.19
TASK	-4.78	SERENE	-2.03
INTERNAL	-4.06	BLEMISH	-2.48
PRECISE	-3.75	COLOSSAL	-3.49
SHRIEK	-4.53	FOE	-3.36
SHRIVEL	-4.33	RESIDE	-2.21
BARRICADE	-3.34	UNSAVORY	-2.71
SEVER	-2.58	ACKNOWLEDGE	-2.03
CONCEITED	-2.58	COMBUSTION	-2.76
WEARY	-3.89	INQUISITIVE	-2.16
BARTER	-3.01	PERPETUAL	-2.87
BRISK	-2.85	PETTY	-1.77
COMMEND	-2.36	RIGID	-3.54
CONSUME	-3.16	PUTRID	-2.08
CONTENT	-3.65	OGRE	-3.18
DEVASTATE	-2.26	PERILOUS	-2.83
EXUBERANT	-2.83	MEAGER	-2.34
DISMAL	-2.90	MONUMENTAL	-1.92
HEX	-4.33	ABHOR	-1.83
INFURIATE	-2.94	DISCLOSE	-2.00
INVINCIBLE	-3.84	VALOR	-2.29
OBSTRUCT	-2.55	AMIABLE	-1.92
POSSESS	-3.94	CLAMOR	-2.26

Appendix E

Sample BIGSCALE Command File
(used for Worksample 741, Form 1)

```
&INST
NAME1=1
NI=110
ITEM1=6
TITLE='M741-1'
MSCDAT='F:M741-1.DAT'
TABLES='00001101100100000001'
IFILE='M741-1.ITM'
PFILE='M741-1.PER'
AFILE='EASY741.ANC'
XWIDE=1
CATEGS=5
CODES='12345'
KEY1='21313225124524553543411245523544131321241515344515512333233
34144341153555335434112455553131142411352133113251 55'
ENDIT=20
&END
TOPIC
SHALLOW
NURSED
GAP
COZY
CABLE
GUIDE
NATURAL
GNAWED
JOURNEY
COMPRESS
GRAVEL
SHRIVEL
GRANTED
FLOCK
CRAM
COMMOTION
DECLARE
```

WITHDRAW
VALUE
PRECISE
APPROPRIATE
GRIEF
GLOBAL
SLAY
FRACTION
BUREAU
INFURIATE
POURED
BRISK
UNIFORM
EMPHASIZE
UNSAVORY
MUDDLE
CONCEITED
CURVATURE
SECURE
DWARF
BLOW
DRAW
POOR
JUMP
LITTLE
FIRE
AIR
YANK
LOVE
CLASS
CHURCH
MARKET
GATE
ABOUT
GROUND
LADY
PLUS
SLIM
FINISH
CAPTAIN
DIVIDE

PIPE
SAVE
NAP
MIDDLE
PATH
FREEZE
APARTMENT
RUN
MUD
DANCE
MAIL
ABOVE
EXPLAIN
EXPLORE
SLIDE
BANK
JOKE
FOLLOW
HANDSOME
CHASE
EARTH
CALL
FEAST
DROP
DARK
PAY
ACROSS
KITCHEN
PLANT
SHOOT
GARDEN
GLOW
ACT
FACE
AGE
WRECK
SADDLE
TURN
JOB
STORY
FIELD

DIZZY
ALARM
BIG
STONE
JOIN
HALL
SLEEPY
BROOM
DRY
BEHIND
END NAMES

Appendix F

Sample BIGSCALE Output
(Worksample 741, Form 1)

```
**********************************************************************
*                                                                    *
*                                                                    *
*                  **** B I G S C A L E ****                         *
*                  ---------------------------------                 *
*                                                                    *
*            - A RASCH PROGRAM FOR RATING SCALE ANALYSIS -           *
*                                                                    *
*                                                                    *
*            PERSON MEASUREMENT, ITEM AND STEP CALIBRATION           *
*                  WITH PERSON AND ITEM FIT ANALYSIS                 *
*                                                                    *
*                                                                    *
*            DIRECT ENQUIRIES TO:                                    *
*                                                                    *
*                  BENJAMIN D. WRIGHT                                *
*                  MESA PRESS                                        *
*                  5835 S KIMBARK AVE                                *
*                  CHICAGO  ILLINOIS 60637                           *
*                                                                    *
*                  (312) 702-1596                                    *
*                  (312) 288-1762                                    *
*                                                                    *
*               COPYRIGHT (C) BENJAMIN D. WRIGHT, 1989               *
* WRITTEN BY BENJAMIN D. WRIGHT, JOHN M. LINACRE, AND MATTHEW SCHULTZ *
*                                                                    *
*                     JANUARY 1990      VERSION 1.53                 *
*                                                                    *
**********************************************************************
```

```
CONTROL VARIABLES
-----------------
 AFILE ='EASY741.ANC
ANCHQU ='n
CATEGS =   5
CHARTF = 0
 CODES ='12345
 DELQU ='n
 DFILE ='
DISTRT = 0
 DSTEP = 0
 ENDIT =  20
FORMAT ='
 IFILE ='M741-1.ITM
INAMES = 0
 ITEM1 =    6
  KEY1 ='21313225124524553543
  KEY2 ='
  KEY3 ='
KEYFRM = 0
KEYSCR ='123
 LCONV =  .0100
 MFIT1 =  2.000
 MFIT2 =  2.000
MISSNG = 255
 MMADJ =  1.00
 MPROX =   4
MSCDAT ='F:M741-1.DAT
 MUCON =  25
 NAME1 =    1
 NCOLS =    0
NEWSCR ='
    NI =  110
OUTFIT = 0
PAFILE ='
PANCHQ ='n
PDELQU ='n
PDFILE ='
 PFILE ='M741-1.PER
 RCONV =  .5000
REALSE = 0
RESCOR ='
RESFRM = 0
 TB11X =   .0
 TB11Y =   .0
  TAB3 =   .0
  TAB4 =   .0
TAB567 =   .0
TABLES ='00001101100100000001
 TITLE ='M741-1
 XFILE ='
 XWIDE = 1
```

ESSENTIAL TABLES	ADDITIONAL TABLES AVAILABLE
3. CATEGORY PROBABILITY CURVES	
4. MOST PROBABLE RESPONSES	
	5. PERSON AND ITEM DISTRIBUTION MAP
	6. ITEM MAP BY NAME
	7. PERSON MAP BY NAME
1. DIAGNOSIS OF MISFITTING PERSONS	2. DIAGNOSIS OF MISFITTING ITEMS
	8. ITEM PLOT OF INFIT VS. DIFFICULTY
	9. ITEM PLOT OF OUTFIT VS. DIFFICULTY
	10. PERSON PLOT OF INFIT VS. ABILITY
	11. PERSON PLOT OF OUTFIT VS. ABILITY
12. ITEM CALIBRATIONS IN ENTRY ORDER	13. ITEM CALIBRATIONS IN DIFFICULTY ORDER
	14. ITEM CALIBRATIONS IN INFIT ORDER
	15. ITEM CALIBRATIONS IN ALPHA ORDER
17. PERSON MEASURES IN ABILITY ORDER	
	16. PERSON MEASURES IN ENTRY ORDER
	18. PERSON MEASURES IN INFIT ORDER
	19. PERSON MEASURES IN ALPHA ORDER
20. PERSON, ITEM AND STEP SUMMARY	

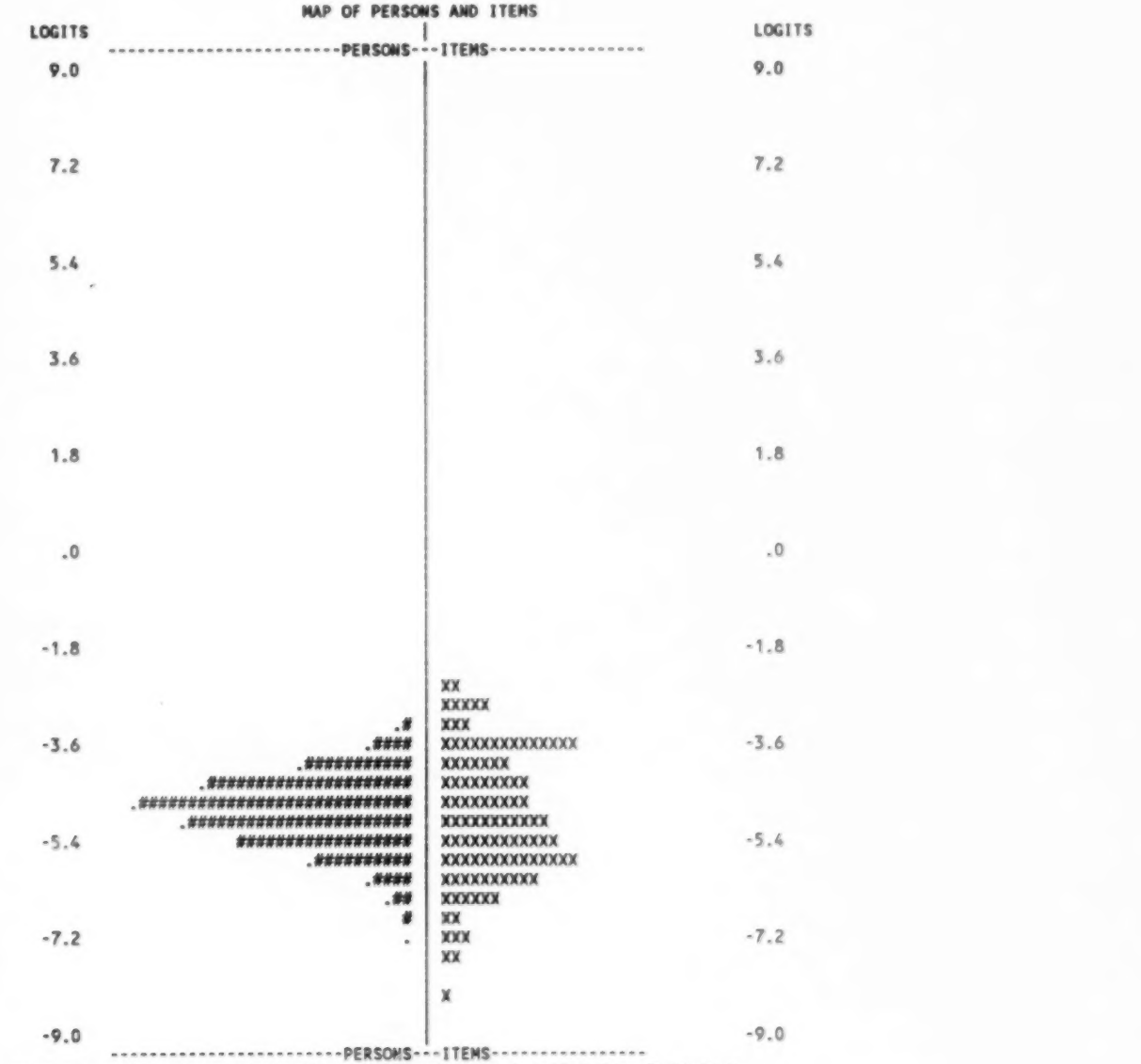
```
TITLE: M741-1                          TIME RUN:   Jan 30  16:32:34  1990
```

CONVERGENCE TABLE

METHOD	ITERATION	MAX LOGIT CHANGE MEASURES	MAX LOGIT CHANGE STEPS	MAX SCORE RESIDUAL MEASURES	MAX SCORE RESIDUAL STEPS
PROX	1	4.7549			
PROX	2	4.4055			
PROX	3	.1866			
PROX	4	.0471			
UCON	1	.0633		3.85	
UCON	2	.0615		1.55	
UCON	3	.0123		1.49	
UCON	4	.0123		.42	
UCON	5	.0077		.97	
UCON	6	.0045		.13	

MAX LOGIT CHANGE = MAXIMUM CHANGE IN ANY LOGIT ESTIMATE
MAX SCORE RESIDUAL = MAXIMUM DISCREPANCY BETWEEN OBSERVED AND EXPECTED SCORES
MEASURES = PERSONS OR ITEMS
STEPS = BETWEEN OBSERVED RESPONSE CATEGORIES

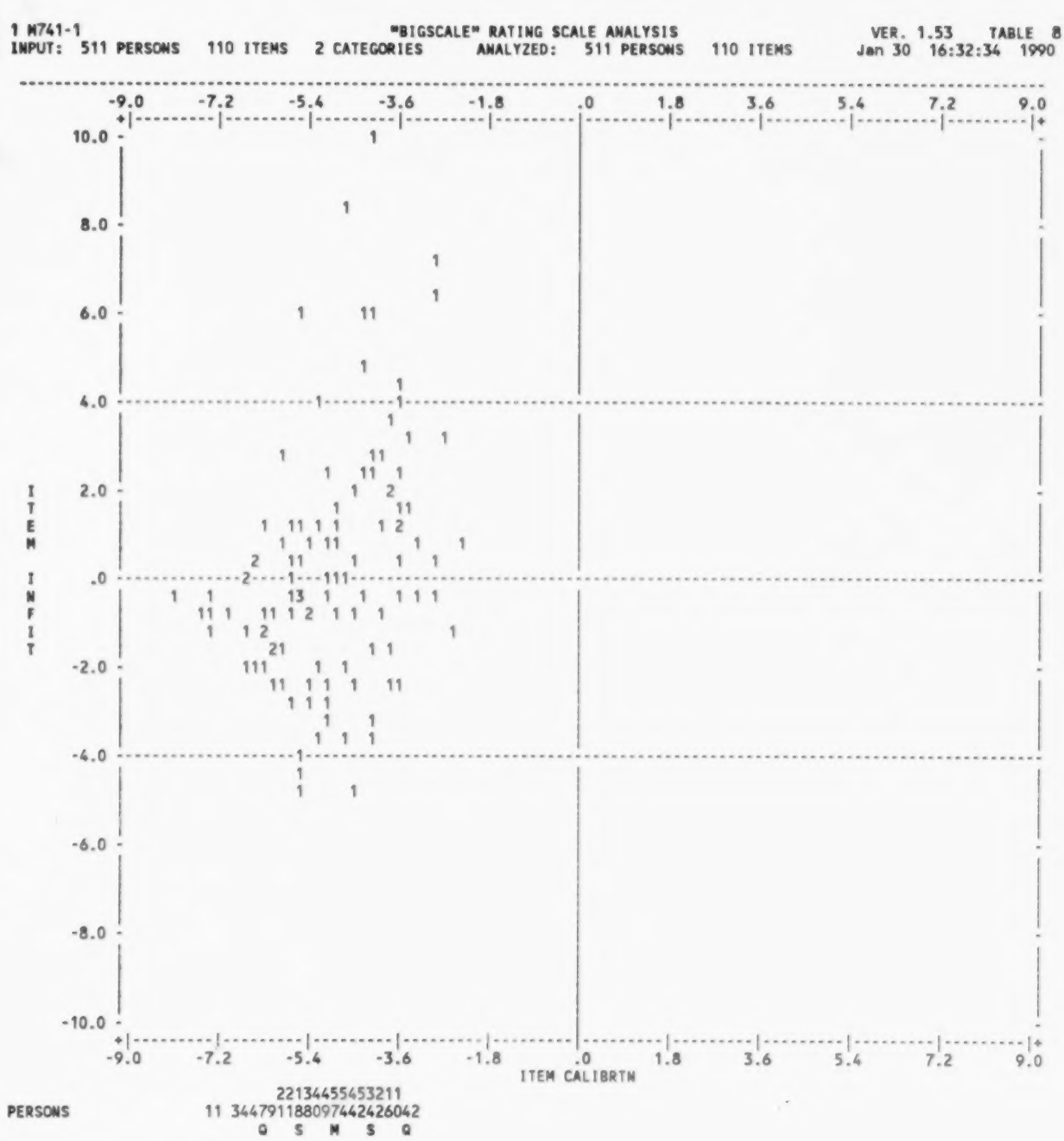
MAP OF PERSONS AND ITEMS

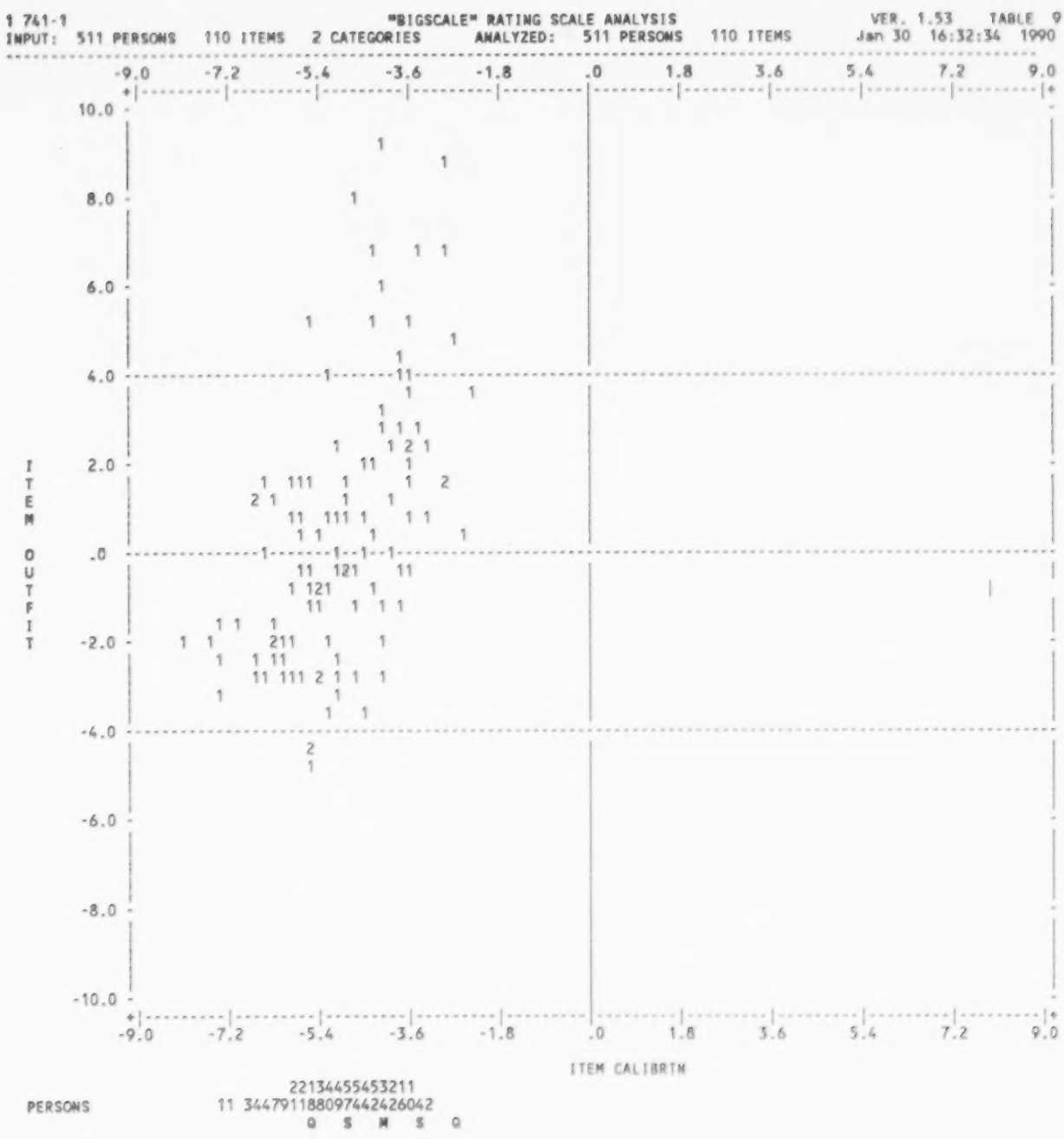


EACH '#' IN THE PERSON COLUMN IS 4 PERSONS; EACH '.' IS 1 TO 3 PERSONS.

MAP OF ITEM+ MEAN STEP

```
STATS |-------------------------------------------------------------------------------------------------------------------
     9.0
      |
      |
      |
```


1 M741-1 "BIGSCALE" RATING SCALE ANALYSIS VER. 1.53 TABLE 8
INPUT: 511 PERSONS 110 ITEMS 2 CATEGORIES ANALYZED: 511 PERSONS 110 ITEMS Jan 30 16:32:34 1990
-9.0 -7.2 -5.4 -3.6 -1.8 .0 1.8 3.6 5.4 7.2 9.0
10.0
8.0
6.0
4.0
2.0
.0
-2.0
-4.0
-6.0
-8.0
-10.0
ITEM
INFIT
ITEM CALIBRTN
PERSONS
22213455453211
11 34479118809744242604 2
Q S M S Q


1 741-1 "BIGSCALE" RATING SCALE ANALYSIS VER. 1.53 TABLE 9
INPUT: 511 PERSONS 110 ITEMS 2 CATEGORIES ANALYZED: 511 PERSONS 110 ITEMS Jan 30 16:32:34 1990
-9.0 -7.2 -5.4 -3.6 -1.8 .0 1.8 3.6 5.4 7.2 9.0
10.0
8.0
6.0
4.0
2.0
.0
-2.0
-4.0
-6.0
-8.0
-10.0
ITEM OUTFIT
ITEM CALIBRTN
PERSONS
22134455453211
11 34479118809744242604 2
Q S M S Q

ITEM+ MEAN STEP STATISTICS -- ENTRY ORDER

NUM	NAME	COUNT	SAMPLE	CALIBRTN	ERROR	MNSQ INFIT		MNSQ OUTFIT		WEIGHT	DISPLACE
[LINKING ITEMS:]											
1	TOPIC	268	509	-5.59A	.10	1.3	6.2	1.3	5.4	.00	-.70
2	SHALLOW	317	510	-5.90A	.10	1.1	2.7	1.1	1.5	.00	-.57
3	NURSED	375	509	-6.02A	.11	1.1	.9	1.1	.8	.00	
4	GAP	174	509	-4.56A	.09	1.0	.4	1.0	.2	.01	-.43
5	COZY	299	510	-5.68A	.10	1.1	1.2	1.0	.7	.00	-.49
6	CABLE	108	508	-4.11A	.10	.9	-3.6	.9	-2.7	.02	-.62
7	GUIDE	276	508	-5.22A	.09	.9	-2.0	.9	-2.1	.01	-.20
8	NATURAL	344	509	-5.65A	.10	1.0	.4	1.0	-.3	.00	
9	GNAWED	222	510	-4.97A	.09	1.0	-.4	1.0	-.2	.01	-.43
10	JOURNEY	267	510	-5.35A	.10	1.0	-1.0	.9	-1.3	.00	-.43
11	COMPRESS	198	510	-4.45A	.09	1.0	-.7	1.0	.6	.01	-.12
12	GRAVEL	327	508	-5.09A	.09	1.0	.7	1.0	.9	.01	.39
13	SHRIVEL	283	507	-4.32A	.09	1.2	4.9	1.2	5.1	.01	.80
14	GRANTED	256	510	-4.30A	.09	1.2	6.0	1.3	6.6	.01	.57
15	FLOCK	206	508	-4.56A	.09	.9	-4.9	.9	-3.8	.01	-.15
16	CRAM	242	509	-4.56A	.09	1.1	2.0	1.1	1.9	.01	.17
17	COMMOTION	207	506	-4.16A	.10	1.1	2.9	1.2	3.2	.02	.27
18	DECLARE	227	506	-4.22A	.10	1.2	5.8	1.3	6.0	.01	.40
19	WITHDRAW	139	506	-4.02A	.10	1.0	-.7	1.0	-.1	.02	-.24
20	VALUE	122	506	-4.15A	.10	.9	-3.4	.9	-1.9	.02	-.52
21	PRECISE	100	507	-3.75A	.10	.9	-1.8	1.0	-.3	.02	-.40
22	APPROPRIATE	108	505	-3.77A	.10	.9	-2.4	.9	-1.4	.02	-.33
23	GRIEF	88	504	-3.61A	.11	.9	-2.3	1.0	-.4	.03	-.41
24	GLOBAL	237	505	-4.72A	.09	1.0	-.1	1.0	-.3	.01	
25	SLAY	143	505	-3.52A	.11	1.3	4.2	1.5	5.1	.03	.33
26	FRACTION	115	504	-3.59A	.11	1.1	1.4	1.3	3.8	.03	
27	BUREAU	266	504	-4.10A	.10	1.4	9.8	1.5	9.0	.02	.93
28	INFURIATE	129	503	-2.94A	.13	1.7	6.5	2.1	6.9	.05	.93
29	POURED	160	504	-3.86A	.10	1.1	2.1	1.2	2.7	.02	.15
30	BRISK	126	504	-2.85A	.13	1.8	7.2	2.5	8.7	.06	1.02
31	UNIFORM	178	508	-3.68A	.10	1.2	3.8	1.2	2.6	.03	.53
32	EMPHASIZE	58	507	-2.83A	.13	1.0	-.5	1.2	1.5	.06	-.17
33	UNSAVORY	79	505	-2.71A	.14	1.4	3.3	1.8	5.0	.07	.34
34	MUDDLE	99	507	-3.33A	.11	1.0	.6	1.3	2.6	.04	
35	CONCEITED	41	506	-2.58A	.14	.9	-1.3	1.1	.5	.08	-.29
36	CURVATURE	216	504	-4.01A	.10	1.1	2.6	1.1	2.5	.02	.55
[TEST ITEMS:]											
37	SECURE	237	507	-4.69	.09	1.0	-1.8	1.0	-1.1	.01	
38	DWARF	393	507	-6.19	.11	.9	-1.5	.8	-2.5	.00	
39	BLOW	124	506	-3.61	.11	1.1	1.5	1.1	1.6	.03	
40	DRAW	68	506	-2.83	.13	1.0	.5	1.2	1.8	.06	
41	POOR	403	504	-6.35	.11	.9	-1.0	.8	-1.9	.00	
42	JUMP	362	505	-5.86	.10	.9	-2.5	.8	-2.9	.00	
43	LITTLE	469	508	-7.52	.17	.9	-.9	.7	-1.9	.00	
44	FIRE	301	502	-5.28	.10	1.0	1.3	1.0	.9	.01	
45	AIR	255	505	-4.85	.09	1.0	1.7	1.1	1.5	.01	
46	YANK	319	506	-5.42	.10	.9	-2.7	.9	-2.8	.00	
47	LOVE	344	506	-5.67	.10	1.0	-.5	1.0	-.8	.00	
48	CLASS	101	507	-3.33	.11	1.2	3.2	1.7	6.6	.04	
49	CHURCH	341	506	-5.63	.10	.9	-3.8	.8	-4.2	.00	
50	MARKET	125	507	-3.62	.11	1.1	2.3	1.3	3.9	.03	
51	GATE	430	505	-6.74	.13	1.0	-.1	1.1	1.2	.00	
52	ABOUT	311	507	-5.34	.10	1.0	-.9	1.0	-.7	.00	
53	GROUND	385	507	-6.10	.11	.9	-2.2	.8	-2.7	.00	
54	LADY	160	506	-3.99	.10	1.0	1.2	1.1	1.1	.02	
55	PLUS	398	506	-6.26	.11	.9	-2.0	.8	-2.3	.00	

NUM	NAME	COUNT	SAMPLE	CALIBRTN	ERROR	MNSQ	INFIT	MNSQ	OUTFIT	WEIGHT	DISPLACE
56	SLIM	327	505	-5.50	.10	.8	-4.3	.8	-4.3	.00	
57	FINISH	422	505	-6.61	.12	.9	-1.9	.7	-2.9	.00	
58	CAPTAIN	137	503	-3.76	.10	1.2	3.6	1.3	4.5	.02	
59	DIVIDE	340	507	-5.62	.10	1.0	-.5	1.0	-.6	.00	
60	PIPE	230	506	-4.63	.09	.9	-3.4	.9	-2.9	.01	
61	SAVE	382	507	-6.06	.11	.9	-1.4	.9	-1.9	.00	
62	MAP	301	506	-5.26	.09	1.1	3.9	1.2	4.0	.01	
63	MIDDLE	350	506	-5.72	.10	.9	-2.9	.8	-3.0	.00	
64	PATH	327	507	-5.49	.10	.8	-4.9	.8	-4.9	.00	
65	FREEZE	335	507	-5.57	.10	1.0	1.1	1.1	1.5	.00	
66	APARTMENT	361	505	-5.85	.10	1.0	-.7	1.0	-.4	.00	
67	RUN	484	506	-8.14	.22	.9	-.4	.6	-1.9	.00	
68	MUD	115	505	-3.51	.11	1.0	-.5	1.1	.7	.03	
69	DANCE	422	507	-6.59	.12	.9	-1.3	.8	-2.4	.00	
70	MAIL	431	506	-6.74	.13	1.0	-.2	1.1	1.1	.00	
71	ABOVE	167	505	-4.07	.10	1.1	2.5	1.1	2.6	.02	
72	EXPLAIN	216	505	-4.52	.09	.9	-2.3	1.0	-.7	.01	
73	EXPLORE	259	506	-4.89	.09	1.0	-.9	1.0	-.5	.01	
74	SLIDE	266	502	-4.96	.09	1.0	-.2	1.0	-.5	.01	
75	BANK	225	505	-4.60	.09	1.3	8.6	1.3	8.2	.01	
76	JOKE	409	506	-6.41	.12	1.0	.3	1.0	-.1	.00	
77	FOLLOW	344	506	-5.67	.10	1.0	-.6	.9	-1.1	.00	
78	HANDSOME	399	506	-6.28	.11	1.1	1.2	1.1	1.3	.00	
79	CHASE	463	505	-7.43	.16	.9	-.5	.8	-1.4	.00	
80	EARTH	299	502	-5.26	.09	.9	-3.6	.9	-3.7	.01	
81	CALL	128	505	-3.66	.11	1.0	.3	1.2	2.1	.03	
82	FEAST	267	503	-4.97	.09	.9	-2.5	.9	-2.3	.01	
83	DROP	359	506	-5.82	.10	1.0	.0	1.0	.4	.00	
84	DARK	313	504	-5.38	.10	.9	-2.3	.9	-2.8	.00	
85	PAY	354	506	-5.77	.10	1.0	-.3	1.0	-.7	.00	
86	ACROSS	272	504	-5.01	.09	1.1	2.4	1.1	2.3	.01	
87	KITCHEN	445	502	-7.06	.14	.9	-.7	.8	-1.7	.00	
88	PLANT	412	504	-6.47	.12	1.0	.3	1.1	1.5	.00	
89	SHOOT	392	503	-6.21	.11	.9	-1.2	.9	-1.9	.00	
90	GARDEN	460	504	-7.37	.16	.9	-.7	.6	-2.4	.00	
91	GLOW	390	498	-6.23	.11	.9	-1.2	.9	-1.5	.00	
92	ACT	184	501	-4.24	.10	1.1	2.2	1.1	1.8	.01	
93	FACE	259	505	-4.89	.09	1.0	.9	1.0	1.0	.01	
94	AGE	96	502	-3.28	.12	1.0	-.5	1.1	.8	.04	
95	WRECK	121	505	-3.58	.11	1.1	1.4	1.2	2.6	.03	
96	SADDLE	318	504	-5.43	.10	1.0	.8	1.0	.5	.00	
97	TURN	391	504	-6.20	.11	1.0	-.7	.9	-.7	.00	
98	JOB	132	503	-3.71	.10	1.1	2.0	1.3	4.1	.02	
99	STORY	187	505	-4.26	.10	1.0	-.2	1.0	.3	.01	
100	FIELD	257	506	-4.87	.09	1.0	.0	1.0	-.3	.01	
101	DIZZY	271	505	-5.00	.09	.9	-3.0	.9	-3.0	.01	
102	ALARM	112	506	-3.47	.11	1.1	1.7	1.3	2.9	.03	
103	BIG	412	505	-6.46	.12	.9	-2.2	.7	-2.9	.00	
104	STONE	460	505	-7.35	.16	.9	-1.2	.5	-3.2	.00	
105	JOIN	265	503	-4.95	.09	.9	-3.2	.9	-2.9	.01	
106	HALL	47	504	-2.41	.15	1.1	.8	1.7	3.8	.09	
107	SLEEPY	175	504	-4.14	.10	.9	-1.6	.9	-1.1	.02	
108	BROOM	256	504	-4.87	.09	1.0	1.3	1.0	1.3	.01	
109	DRY	351	503	-5.75	.10	1.0	.6	1.1	1.6	.00	
110	BEHIND	372	505	-5.96	.11	.9	-1.6	.9	-1.9	.00	

"WEIGHT"S ARE MULTIPLICATIVE ON A RATIO SCALE. "CALIBRTN"S ARE ADDITIVE ON AN INTERVAL SCALE.
THE STANDARD ERROR OF A WEIGHT IS THE VALUE OF THE "WEIGHT" TIMES THE VALUE OF THE CALIBRTN "ERROR"

SUMMARY OF 511 MEASURED PERSONS

	COUNT	TEST	MEASURE	ERROR	MNSQ	INFIT	MNSQ	OUTFIT
MEAN	56.7	108.8	-4.88	.23	1.0	.1	1.1	.2
S.D.	14.3	7.5	.70	.02	.2	1.4	.3	1.4

RMSE	.23	ADJ.S.D.	.66	PERSON SEP	2.89	PERSON SEP REL. .89

SUMMARY OF 110 CALIBRATED ITEMS CENTERED ON MEAN STEP VALUE

	COUNT	SAMPLE	CALIBRTN	ERROR	MNSQ	INFIT	MNSQ	OUTFIT
MEAN	263.6	505.6	-4.96	.11	1.0	.2	1.1	.5
S.D.	115.7	2.1	1.22	.02	.1	2.7	.3	2.9

RMSE	.11	ADJ.S.D.	1.22	ITEM SEP	11.14	ITEM SEP REL. .99

SUMMARY OF CALIBRATED STEPS

LABEL	VALUE	COUNT	MEASURE	ERROR	RESIDUAL
0	0	26622	NONE		3.4
1	1	28995	NONE		-3.4

OUTFIT: MEAN SQUARE STANDARD RESIDUAL -- STANDARDIZED TO (0,1) EXPECTATION
INFIT: MEAN SQUARE INFORMATION RESIDUAL -- STANDARDIZED TO (0,1) EXPECTATION
SEPARATION: RATIO OF ADJUSTED SD TO ROOT MEAN SQUARE ERROR
RELIABILITY: RATIO OF ADJUSTED VARIANCE TO OBSERVED VARIANCE

Appendix G

Database Structure

Definitions of Structure Terms:

Field	Nth variable in the database record
Field Name	Name of the field
Type	The type of field. "Character" includes alphanumeric data; "Numeric" includes only numbers; "Logical" values are "T" for True or "F" for False; "Memo" contains unlimited alphanumeric data in a word-processing format.
Width	The numbers of places held for data within that field. Logical fields always contain 1 space. Memo fields are listed as containing 10 spaces but are actually variable, depending on the number of characters entered.
Dec	For numeric fields, this is the number of places to the right of the decimal place. For other fields, this is irrelevant.

CAT1 CAT2 CAT3 CAT4 CAT5	The type of choice of the first through fifth item choice. For the most part, the synonym is marked as "s," and the other choices not marked. Beginning with the Worksample 735 test series, all choice types are identified (synonym, sound-alike, close mislead, same situation, antonym).
CHOICE1 CHOICE2 CHOICE3 CHOICE4 CHOICE5	The text of the five item choices.
CURRENT	Whether the given item administration is the administration that was ultimately used to estimate VSS or was superseded by a later administration (for each meaning of a word). Current is only true once for each meaning of a word, even though the word may have been used in several different items, each of which may have been administered on multiple occasions.
DIFF	A difficulty rating of the word on the scale of 1 to 5 originally devised by Gary Supanich to subjectively estimate the difficulty of the word. The use of *The Living Word Vocabulary* (Dale & O'Rourke, 1981) makes this value obsolete.
DISCUSSION	*Wordbook* discussion.
ERROR	The standard error of the item's logit measure.
EX1	*Wordbook* Exercise 1.
EX1ANS	*Wordbook* Exercise 1 answer.
EX2	*Wordbook* Exercise 2.
EX2ANS	*Wordbook* Exercise 2 answer.

EX3A	*Wordbook* Exercise 3 Choice A.
EX3B	*Wordbook* Exercise 3 Choice B.
EX3C	*Wordbook* Exercise 3 Choice C.
EX3ANS	*Wordbook* Exercise 3 answer.
FORM1 FORM2 FORM3 FORM4 FORM5	First through fifth *Wordbook* alternative form of the word.
GLOBMEAS	The VSS value of the word expressed in logits (50% correct).
GOODSAMPLE	Whether the overall ability of the sample was appropriate for the difficulty level of the item. (See Research Memorandum 1990-2 for complete details.)
GOODQUAL	Whether the item is of good quality relative to the values of INFIT, OUTFIT, and their related mean squares. (See Research Memorandum 1990-1 for details.)
INFIT	The INFIT value.
ITEM	The item number of the item within the test.
ITEMWORD	The tested word.
LINK_ANCHR	Whether the item is a linking item, an equating item, or neither.
MEANING	A now-obsolete code that refers to the numerical code of the meaning tested by the Foundation (the new system will list the word and its synonym).
MEANSQ	The mean square INFIT value.

MEASURE	The logit measure of the item relative to the sample, not necessarily anchored to Worksample 705-1. New items are automatically anchored, and so the MEASURE field is the same as GLOBMEAS.
OUTFIT	The OUTFIT value.
PARTSPEECH	A one-character code giving the word's part of speech (n=noun; v=verb; a=adjective).
PHRASE	The phrase in which the word was tested. Our testing program no longer uses phrases.
REVIEW	*Wordbook* review item.
REVIEWANS	*Wordbook* review item answer.
REVISION	The number of the revision of a given word for the given meaning (e.g., 1st revision, 2nd revision, and so on).
ROOTWORD	The ITEMWORD stripped of suffixes and prefixes.
SAMPLE	The number of people who took the item.
SCORE	The number of persons who answered the item correctly.
SPEECH1 SPEECH2 SPEECH3 SPEECH4 SPEECH5	Part of speech of the first through fifth alternative forms of the word.
TEST	The worksample number for the test item (e.g., "705-1").
VSS80	The VSS value of the word.
WEIGHT	The statistical "weight" of the item.

Structure of ITEMS.DBF

Number of data records: 6943
Date of last update: 06/08/89

Field	Field Name	Type	Width	Dec
1	ITEMWORD	Character	20	
2	MEANING	Character	1	
3	DIFF	Numeric	1	
4	REVISION	Numeric	1	
5	ROOTWORD	Character	20	
6	PARTSPEECH	Character	1	
7	PHRASE	Character	45	
8	CHOICE1	Character	20	
9	CAT1	Character	1	
10	CHOICE2	Character	20	
11	CAT2	Character	1	
12	CHOICE3	Character	20	
13	CAT3	Character	1	
14	CHOICE4	Character	20	
15	CAT4	Character	1	
16	CHOICE5	Character	20	
17	CAT5	Character	1	
**Total **			195	

Structure of USED.DBF

Number of data records: 4002
Date of last update: 11/24/87

Field	Field Name	Type	Width	Dec
1	TEST	Character	6	
2	ITEM	Numeric	3	
3	ITEMNAME	Character	20	
** Total **			30	

Structure of DISCUSS.DBF

Number of data records: 1000
Date of last update: 11/20/89

Field	Field Name	Type	Width	Dec
1	ROOTWORD	Character	25	
2	MEANING	Character	1	
3	EX1	Character	75	
4	EX1ANS	Logical	1	
5	EX2	Character	150	
6	EX2ANS	Logical	1	
7	EX3A	Character	25	
8	EX3B	Character	25	
9	EX3C	Character	25	
10	EX3ANS	Character	1	
11	REVIEW	Character	200	
12	REVIEWANS	Character	25	
13	DISCUSSION	Memo	10	
14	FORM1	Character	25	
15	SPEECH1	Character	8	
16	FORM2	Character	25	
17	SPEECH2	Character	8	
18	FORM3	Character	25	
19	SPEECH3	Character	8	
20	FORM4	Character	25	
21	SPEECH4	Character	8	
22	FORM5	Character	25	
23	SPEECH5	Character	8	
**Total **			730	

Structure of ALLSTATS.DBF

Number of data records: 6257
Date of last update: 03/06/90

Field	Field Name	Type	Width	Dec
1	WORD	Character	16	
2	TEST	Character	6	
3	ITEM	Numeric	3	
4	VSS80	Numeric	4	
5	GLOBMEAS	Numeric	5	2
6	SCORE	Numeric	3	
7	SAMPLE	Numeric	3	
8	WEIGHT	Numeric	4	2
9	MEASURE	Numeric	5	2
10	ERROR	Numeric	4	2
11	MEANSQ	Numeric	3	1
12	OUTFIT	Numeric	4	2
13	INFIT	Numeric	4	2
14	LINK_ANCHR	Logical	1	
15	GOODSAMPLE	Logical	1	
16	GOODQUAL	Logical	1	
17	CURRENT	Logical	1	
18	TEMP	Logical	1	
** Total **			70	

Structure of STATS.DBF

Number of data records: 8761
Date of last update : 06/22/88

Field	Field Name	Type	Width	Dec
1	ITEMWORD	Character	25	
2	REVISION	Numeric	1	
3	TEST	Character	6	
4	ITEM	Numeric	3	
5	LINKING	Logical	1	
6	WORDBOOK	Numeric	2	
7	VSS80	Numeric	3	
**Total **			42	

Appendix H

Test Series Contained in Each Database

Test Series Contained in USED.DBF

690A
690B
690C
705-1
705-10
705-11
705-2
705-3
705-4
705-5
705-6
705-7
705-8
705-9
722A
722B
722C
722D
722E
722F
722G
722H
722I
722J
735A
735B
738 series
Wordbook 1
Wordbook 2
Wordbook 3
Wordbook 4
Wordbook 5
Wordbook 6
Wordbook 7
Wordbook 8

Test Series Contained in ALLSTATS.DBF

690A
690B
690C
704
705-1
705-10
705-11
705-2
705-3
705-4
705-5
705-6
705-7
705-8
705-9
708A
708B
708C
722A
722B
722C
722D
722E
722F
722G
722H
722I
722J
734A*
735A
735B
735C
735D
735E
738A
738B
738C
738D
738E
738F
738G
738H
738I
738J
738K
738L
738M
Wordbook 1
Wordbook 2
Wordbook 3
Wordbook 4
Wordbook 5
Wordbook 6
Wordbook 7
Wordbook 8

Test Series Contained in STATS.DBF

176AB
176AD
176BA
180AD
180AE
180AF
180BA
180BB
271B
271C
600-A
600AA
600AB
600C
603F
604-E
605-BA
605-CA
605-FA
620D
620E
620F
629A
629AA
629AC
629B
641CA
641CC
641DC
641EA
641EC
641F
641G
649A
678A
680A
680B
684A
687
687B
698A
699-1
699-10
699-11
699-12
699-13
699-14
699-15
699-16
699-2
699-3
699-4
699-5
699-6
699-7
699-8
699-9
702
704
708A
708B
708C
734A*
95AD
95BC
95CC
95DB
95EA
95GA
95H
95I
95JB
GINNB
GINNC
GINNF

References

Andersen, E. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, 32*, 283-301.

Andersen, E. (1973). Conditional inference for multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26*, 283-301.

Andersen, E. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69-81.

Bowker, R. (1979a). *Wordbooks 1-6*. New York: Johnson O'Connor Research Foundation.

Bowker, R. (1979b). *Wordbook teacher's manual*. Boston: Johnson O'Connor Research Foundation.

Bowker, R. (1982). Producing new *Wordbooks*. *The Research Newsletter, 2* (5), 3-5.

Bowker, R. (1983). *Wordbooks 7-8*. New York: Johnson O'Connor Research Foundation.

Brogden, H. (1977). The Rasch model, the law of comparative judgement and additive conjoint measurement. *Psychometrika, 37*, 29-51.

Choppin, B. (1976). Recent developments in item banking. In *Advances in psychological and educational measurement*. New York: Wiley.

Choppin, B. (1978). *Item banking and the monitoring of achievement*. Slough: National Foundation for Educational Research in England and Wales.

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary*. Chicago: World Book--Child Craft International, Inc..

English Vocabulary manual. (1981). R. Bowker. Boston: Johnson O'Connor Research Foundation.

Gershon, R., & Schroeder, D. (1987, April). *Building a vocabulary item bank: Some findings*. Paper presented at the International Objective Measurement Workshop, Chicago.

Kelderman, H. (1986). *Common item equating using the loglinear Rasch model* (Research Report 86-9). Enschede, Netherlands: University of Twente, Department of Education, Division of Educational Measurement and Data Analysis.

Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement, 21*, 315-330.

Mislevy, R. (1990). Foundations of a new test theory. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedogogiske Institut (Chicago: University of Chicago Press, 1980).

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathemetical Statistics and Probability*, 321-333.

Rasch, G. (1966a). An individual approach to item analysis. In P. F. Lazarsky & N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 89-108). Chicago: Science Research Associates.

Rasch, G. (1966b). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Rasch, G. (1968). A mathematical theory of objectivity and its consequences for model construction. In *Report from the European Meeting on Statistics, Econometrics and Management Sciences*, Amsterdam, 1968.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Research Memorandum 1990-2. *Description of the Worksample 741 mislead analysis*. R. Gershon. Chicago: Johnson O'Connor Research Foundation.

Schultz, M. (1988). A Rasch program for one-step item banking. *Rasch Measurement SIG Newsletter, 1*(2), 4-5.

Smith, R. (1984). Person fit in the Rasch model. *Applied Psychological Measurement, 46*, 359-372.

Statistical Bulletin 1980-33. *Conversions from Wks. 690 B, 690 C to Vocabulary Scale Scores; Subtest assignments using the Vocabulary Placement Test, Wks. 695 A.* R. Bowker. Boston: Johnson O'Connor Research Foundation.

Technical Report 1988-3. *Index of words in the Johnson O'Connor Research Foundation, Inc. vocabulary item bank.* R. Gershon. Chicago: Johnson O'Connor Research Foundation.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

Wright, B. D. (1977a). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

Wright, B. D. (1977b). Misunderstanding the Rasch model. *Journal of Educational Measurement, 14*, 219-255.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why and how. *Journal of Educational Measurement, 21*, 331-345.

Wright, B. D., Congdon, R., & Rossner, M. (1987). *MSCALE*. Chicago: MESA Press.

Wright, B. D., Linacre, M., & Schultz, M. (1989). *BIGSCALE*. Chicago: MESA Press.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items with the Rasch model* (Research Memorandum No. 23C). Chicago, IL: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

JOHNSON O'CONNOR RESEARCH FOUNDATION
HUMAN ENGINEERING LABORATORY